1.- Se ha encuestado a 100 familias en una ciudad sobre su gasto mensual en ocio (Y), y sus ingresos mensuales (X). En la siguiente tabla se presentan los resultados obtenidos, donde la variable X viene expresada en miles de unidades monetarias, y la variable Y en unidades monetarias.

ingresos mensuales (X)	gasto mensual en ocio (Y)			
	0-2000	2000-8000	8000-20000	20000-80000
60-100	4	1	1	-
100-150	9	8	3	-
150-200	9	12	20	3
200-300	5	8	12	3
300-500	1	1	-	-

Se pide:

a) Obtenga el ingreso mensual medio de estas 100 familias.
b) Calcule el índice de Gini de la distribución de ingresos de las familias cuyo gasto en ocio es superior a las 8.000 unidades monetarias.
c) Discuta cuál de las dos distribuciones marginales es más homogénea.
d) Razone si X e Y son independientes estadísticamente.
e) Obtenga qué valores determinan el gasto en ocio del 50% central de las familias cuyos ingresos no superan las 200.000 unidades monetarias.

a) y c)

Ingresos mensuales						
Datos cuantitativos continuos						
Valores			Frecuencias			
L.inf.	L.sup	x_i	n_i	n_i*x_i	x_i^2	$n_i*x'_i$^2
60	100	80	6	480	6400	38400
100	150	125	20	2500	15625	312500
150	200	175	44	7700	30625	1347500
200	300	250	28	7000	62500	1750000
300	500	400	2	800	160000	320000
Total			100	18480		3768400

$$\bar{x}_I = \frac{\sum n_i x_i'}{N} = \frac{18480}{100} = 184{,}8$$

$$s_I^2 = \frac{\sum n_i x_i'^2}{N} - \bar{x}_I^2 = \frac{3768400}{100} - 184{,}8^2 = 3532{,}96$$

$$s_I = \sqrt{3532{,}96} = 59{,}43$$

$$CV_I = \frac{s_I}{\bar{x}_I} = \frac{59{,}43}{184{,}8} = 0{,}32$$

Gasto mensual en ocio						
Datos cuantitativos continuos						
Valores			Frecuencias			
L.inf.	L.sup	x_i	n_i	n_i*x_i	x_i^2	n_i*x_i^2
0	2000	1000	28	28000	1000000	28000000
2000	8000	5000	30	150000	25000000	750000000
8000	20000	14000	36	504000	196000000	7056000000
20000	80000	50000	6	300000	2500000000	15000000000
Total			100	982000		22834000000

$$\bar{x}_O = \frac{\sum n_i x_i'}{N} = \frac{982000}{100} = 9820$$

$$s_O^2 = \frac{\sum n_i x_i'^2}{N} - \bar{x}_O^2 = \frac{22834000000}{100} - 9820^2 = 131907600$$

$$s_I = \sqrt{131907600} = 11485{,}10$$

$$CV_O = \frac{s_O}{\bar{x}_O} = \frac{11485{,}10}{9820} = 1{,}17$$

Como $CV_I < CV_O$ es más homogénea la distribución de los ingresos, que la distribución del gasto en ocio.

b)

x_i	n_i	N_i	p_i	x_i*n_i	$(xi*ni)\uparrow$	q_i
80	1	1	0,02	80	80	0,009720535
125	3	4	0,1	375	455	0,055285541
175	23	27	0,64	4025	4480	0,544349939
250	15	42		3750	8230	
400	0	42		0	8230	
Total	42		0,76	8230	8230	0,609356015

El índice de concentración de Gini es

$$: I_G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{0{,}61}{0{,}76} = 0{,}20$$

d) Para que ambas variables sean independientes tiene que cumplirse:

$f_{ij} = f_{i.} \cdot f_{.j} \forall i, j$ Como $f_{14} = 0$ y $f_{1.} \cdot f_{.4} = \frac{6}{100}\frac{6}{100} \neq 0$

$f_{1.} \cdot f_{.4} \neq f_{14}$ y por tanto no son independientes.

e)

Gasto mensual en ocio			
Datos cuantitativos continuos			
Valores		Frecuencias	
L.inf.	L.sup	n_i	N_i
0	2000	22	22
2000	8000	21	43
8000	20000	24	67
20000	80000	3	70
Total		70	

$$\frac{N}{4} = \frac{70}{4} = 17{,}5 \Rightarrow Q_1 \in [0{,}2000)$$

$$\left.\begin{array}{l} 22 \_\_\_\_\_ 2000 \\ 17{,}5 \_\_\_\_\_ x \end{array}\right\} x = \frac{2000.17{,}5}{22} = 1590{,}91$$

$$Q_! = 0 + 1590{,}91 = 1590{,}91$$

$$\frac{3N}{4} = \frac{3.70}{4} = 52{,}5 \Rightarrow Q_3 \in [8000{,}20000)$$

$$\left.\begin{array}{l} 24 \_\_\_\_\_ 12000 \\ 9{,}5 \_\_\_\_\_ x \end{array}\right\} x = \frac{12000.9{,}5}{24} = 4750$$

$$Q_{3!} = 8000 + 4750 = 12750$$

PROBLEMA Nº 2

De una distribución bidimensional (X, Y) se conocen las siguientes distribuciones de frecuencias:

X	$n_{i\cdot}$
1	10
2	20
3	15

Y/X=1	$f_{Y/X=1}$
2	0,2
4	0,5
6	0,3

Y/X=2	$f_{Y/X=2}$
2	0,45
4	0,30
6	0,25

Y/X=3	$f_{Y/X=3}$
2	0,4
4	0,6
6	0

a) Determinar la distribución de frecuencias absolutas conjuntas de (X,Y), la distribución marginal de Y y la condicionada de X al valor de Y = 2

b) Calcular la media y la varianza de la variable X condicionada a que Y = 2 y la de la variable 2X - 1 condicionada a que Y = 2

SOLUCIÓN

X \ Y	2	4	6	$n_{i\bullet}$
1				10
2				20
3				15
$n_{\bullet j}$				N = 45

$$f_{j/i} = \frac{n_{ij}}{n_{i\bullet}} \Rightarrow n_{ij} = n_{i\bullet} f_{j/i}$$

Y/X=1	$f_{Y/X=1}$
2	0,2
4	0,5
6	0,3
	1

$$n_{11} = n_{1\bullet} f_{1/1} = 10 \cdot 0{,}2 = 2$$
$$n_{12} = n_{1\bullet} f_{2/1} = 10 \cdot 0{,}5 = 5$$
$$n_{13} = n_{1\bullet} f_{3/1} = 10 \cdot 0{,}3 = 3$$

Y/X=2	$f_{Y/X=2}$
2	0,45
4	0,30
6	0,25
	1

$$n_{21} = n_{2\bullet} f_{1/2} = 20 \cdot 0{,}45 = 9$$
$$n_{22} = n_{2\bullet} f_{2/2} = 20 \cdot 0{,}3 = 6$$
$$n_{23} = n_{2\bullet} f_{3/2} = 20 \cdot 0{,}25 = 5$$

Y/X=3	$f_{Y/X=3}$
2	0,4
4	0,6
6	0
	1

$$n_{31} = n_{3\bullet} f_{1/3} = 15 \cdot 0{,}4 = 6$$
$$n_{32} = n_{3\bullet} f_{2/3} = 15 \cdot 0{,}6 = 9$$
$$n_{33} = n_{3\bullet} f_{3/3} = 15 \cdot 0 = 0$$

Luego, la distribución de frecuencias absolutas conjuntas de (X,Y) es:

Y / X	*2*	*4*	*6*	$n_{i\bullet}$
1	*2*	*5*	*3*	*10*
2	*9*	*6*	*5*	*20*
3	*6*	*9*	*0*	*15*
$n_{\bullet j}$	*17*	*20*	*8*	$N = 45$

La distribución marginal de Y es:

Y	$n_{\bullet j}$
2	17
4	20
6	8

La distribución de $(X/Y=2)$ es:

$X/Y=2$	$n_{(X/Y=2)}$	$f_{(X/Y=2)}$
1	2	2/17
2	9	9/17
3	6	6/17
	17	1

b)

$X/Y=2$	n_i	$x_i n_i$	$x_i^2 n_i$
1	2	2	2
2	9	18	36
3	6	18	54
	17	38	92

Media de $(X/Y=2) = \dfrac{38}{17} = 2{,}2353$

Varianza de $(X/Y=2) = \dfrac{92}{17} - \left(\dfrac{38}{17}\right)^2 = 0{,}4152$

Media de $(2X-1/Y=2) = 2 \cdot \dfrac{38}{17} - 1 = 2 \cdot 2{,}2353 - 1 = 3{,}4706$

Varianza de $(2X-1/Y=2) = 2^2 \cdot 0{,}4152 = 1{,}6608$

3.- Se han estudiado 50 empresas que invertían por primera vez en publicidad, observándose para cada una de ellas los beneficios obtenidos en millones de euros (Y) y el capital invertido en miles de euros (X):

X	Y			
	[0,20]	**(20,50]**	**(50,100]**	**(100,200]**
[2,4]	13	2	2	0
(4,10]	0	1	10	0
(10,15]	0	0	3	8
(15,20]	0	0	1	10

Se pide:

a) ¿Cuál de las dos distribuciones resulta más homogénea en torno a su media? ¿Por qué?
b) Calcular la inversión en publicidad más frecuente para las empresas cuyo beneficio se sitúa entre 50 y 100 millones de euros.
c) ¿Qué volúmenes de inversión en publicidad delimitan el 80% central de las empresas que han obtenido un beneficio entre 50 y 100 millones de euros?
d) ¿Son ambas variables estadísticamente independientes?
e) Ajuste una recta de regresión que permita predecir los beneficios en función del capital invertido en publicidad.
f) ¿Qué beneficios se esperan si se invierte en publicidad 8.000 €?
g) Estudie la bondad del ajuste

X : "Capital invertido en publicidad (en miles de €)"

Y : "Beneficios obtenidos (en millones de €)"

y_j / x_i		*[0, 20]*	*(20, 50]*	*(50, 100]*	*(100, 200]*	$n_{i\bullet}$
		10	*35*	*75*	*150*	
[2, 4]	*3*	*13*	*2*	*2*	*0*	*17*
(4, 10]	*7*	*0*	*1*	*10*	*0*	*11*
(10, 15]	*12,5*	*0*	*0*	*3*	*8*	*11*
(15, 20]	*17,5*	*0*	*0*	*1*	*10*	*11*
$n_{\bullet j}$		*13*	*3*	*16*	*18*	*50*

a) Para estudiar la homogeneidad de las distribuciones calculamos los coeficientes de variación marginales de X y de Y:

x_i	$n_{j\bullet}$	$x_i n_{i\bullet}$	$x_i^2 n_{i\bullet}$
3	17	51	153
7	11	77	539
12,5	11	137,5	1718,75
17,5	11	192,5	3368,75
	50	458	5779,5

$$\bar{x} = \frac{\sum x_i n_{i\bullet}}{N} = \frac{458}{50} = 9{,}16 \text{ miles de } €$$

$$S_x^2 = \frac{\sum x_i^2 n_{i\bullet}}{N} - \bar{x}^2 = \frac{5779{,}5}{50} - 9{,}16^2 = 31{,}6844$$

$$S_x = \sqrt{31{,}6844} = 5{,}6289 \text{ miles de } €$$

$$CV_x = \frac{S_x}{\bar{x}} = \frac{5{,}6289}{9{,}16} = 0{,}6145$$

y_j	$n_{\bullet j}$	$y_j n_{\bullet j}$	$y_j^2 n_{\bullet j}$
10	13	130	1300
35	3	105	3675
75	16	1200	90000
150	18	2700	405000
	50	4135	499975

$$\bar{y} = \frac{\sum y_j n_{\bullet j}}{N} = \frac{4135}{50} = 82{,}7 \text{ millones de } €$$

$$S_y^2 = \frac{\sum y_j^2 n_{\bullet j}}{N} - \bar{y}^2 = \frac{499975}{50} - 82{,}7^2 = 3160{,}21$$

$$S_y = \sqrt{3160{,}21} = 56{,}2157 \text{ millones de } €$$

$$CV_y = \frac{S_y}{\bar{y}} = \frac{56{,}2157}{82{,}7} = 0{,}6798$$

Como $CV_x < CV_y$ la distribución marginal de X es un poco más homogénea que la de Y.

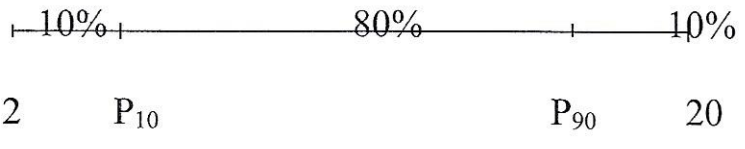
b) Nos piden la MODA de la distribución condicionada $(X / Y \in (50, 100])$

$x / y \in (50,100]$	n_i	c_i	$h_i = n_i / c_i$	$N_i\uparrow$
[2, 4]	2	2	1	2
(4, 10]	10	6	1,67	12
(10, 15]	3	5	0,6	15
(15, 20]	1	5	0,2	16
	16			

Intervalo modal = [4, 10)

$$Mo = \frac{4+10}{2} = 7 \text{ mil } €$$

c) Seguimos con la variable $(X / Y \in (50, 100])$



1º) Calculamos $\frac{kN}{q} = \frac{10N}{100} = \frac{10 \cdot 16}{100} = 1{,}6$

2º) El primer $N_i^{\uparrow} \geq 1{,}6$ corresponde al intervalo [2, 4)

3º) $$P_{10} = L_{i-1} + \frac{\frac{kN}{q} - N_{i-1}^{\uparrow}}{n_i} \cdot c_i = 2 + \frac{1{,}6 - 0}{2} \cdot 2 = 3{,}6$$

1º) Calculamos $\frac{kN}{q} = \frac{90N}{100} = \frac{90 \cdot 16}{100} = 14{,}4$

2º) El primer $N_i^{\uparrow} \geq 14{,}4$ corresponde al intervalo [10, 15)

3º) $$P_{90} = 10 + \frac{14{,}4 - 12}{3} \cdot 5 = 14$$

d) ¿ X, Y independientes ?

$$X \text{ e } Y \text{ indep.} \iff f_{ij} = f_{i\bullet} \cdot f_{\bullet j}, \ \forall i, j$$

Como, por ejemplo, para i = 2, j = 1 $\quad 0 \neq \frac{11 \cdot 13}{50}$

las variables son dependientes.

e) ¿ Recta de regresión de Y sobre X ?

¿ S_{xy} ?

$$S_{xy} = \frac{\sum_i \sum_j x_i y_j n_{ij}}{N} - \bar{x} \cdot \bar{y} = \frac{51920}{50} - 757{,}532 = 280{,}868$$

	10	35	75	150
3	13	2	2	0
7	0	1	10	0
12,5	0	0	3	8
17,5	0	0	1	10

$\rightarrow$

390	210	450	0
0	245	5250	0
0	0	2812,5	15000
0	0	1312,5	26250

1050
5495
17812,5
27562,5
51920

$$y-\bar{y}=\frac{S_{xy}}{S_x^2}(x-\bar{x}) \leftrightarrow y-82,7=\frac{280,87}{31,68}(x-9,16) \leftrightarrow y=8,87x+1,49$$

f) ¿ $\hat{y}(8)$?

$$\hat{y}(8)=8,87\cdot 8+1,49=72,45$$

g) $$R^2=\frac{S_{xy}^2}{S_x^2\cdot S_y^2}=\frac{280,87^2}{31,68\cdot 3160,21}=0,7880$$

4.- En una determinada región se observó el precio del trigo (en euros/kg) y la cantidad producida (en miles de toneladas) durante algunos años, obteniéndose la siguiente tabla:

Precio	35	31	42	60	52	49	61	50	55	58
Cantidad	100	140	120	110	200	200	110	160	160	200

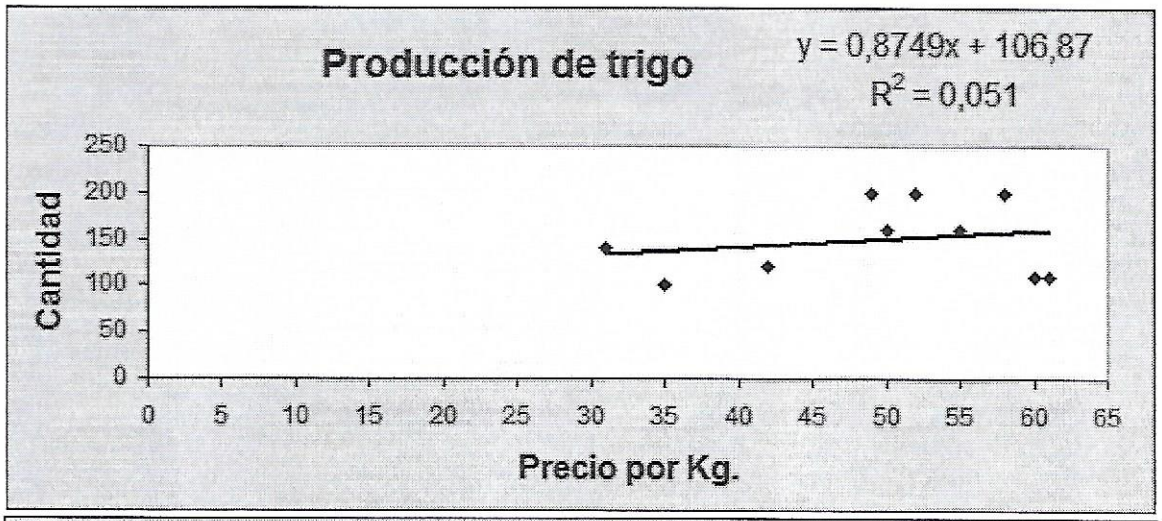
Se pide:

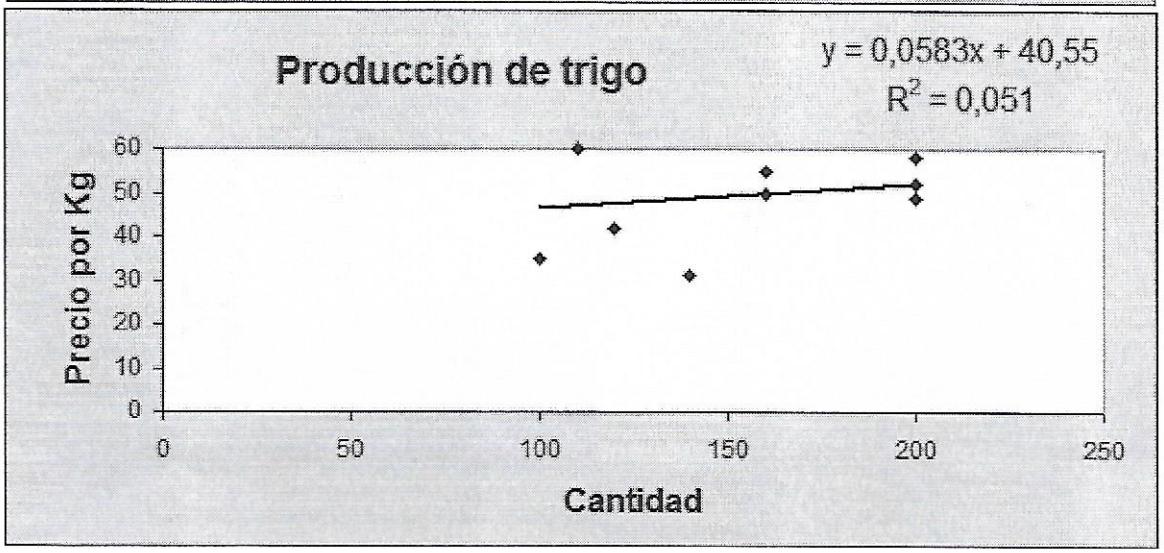
a) Con un precio de 50€, ¿cuál sería la cantidad de producción esperada según una regresión lineal mínimo-cuadrática?. ¿Qué fiabilidad presenta dicha predicción?

b) ¿En cuánto se incrementaría la producción si el precio aumentase en 2 €/Kg.

c) Exprese la varianza total como descomposición de varianza no explicada y varianza explicada.

Producción de trigo					
Variable estadística bidimensional					
Años	Precio	Cantidad			
	x_i	y_i	x_i^2	y_i^2	x_i*y_i
1	35	100	1225	10000	3500
2	31	140	961	19600	4340
3	42	120	1764	14400	5040
4	60	110	3600	12100	6600
5	52	200	2704	40000	10400
6	49	200	2401	40000	9800
7	61	110	3721	12100	6710
8	50	160	2500	25600	8000
9	55	160	3025	25600	8800
10	58	200	3364	40000	11600
Total	493	1500	25265	239400	74790
Medias	49,3	150			
Varianzas	96,01	1440			
Covarianza	84				
D.típicas	9,7984693	37,94733192			
R y R^2	0,2259123	0,05103635			
s_{xy}/s_x^2	0,875		106,8670		
sxy/s_y^2	0,0583		40,55		
Recta Y/X	Y=0,8749X+106,87				
50	150,612				
Recta X/Y	X=0,0583Y+40,55				
100	46,383				





a) Hacemos los cálculos oportunos para obtener la ecuación de la recta de regresión de Y/X:

$$\bar{\bar{x}} = \frac{\sum n_i x_i'}{N} = \frac{493}{100} = 49{,}3$$

$$s_x^2 = \frac{\sum n_i x_i'^2}{N} - \bar{\bar{x}}^2 = \frac{25265}{10} - 49{,}3^2 = 96{,}01$$

$$s_x = \sqrt{96{,}01} = 9{,}7985$$

$$\bar{\bar{y}} = \frac{\sum n_i y_i'}{N} = \frac{1500}{10} = 150$$

$$s_y^2 = \frac{\sum n_i y_i'^2}{N} - \bar{\bar{y}}^2 = \frac{239400}{10} - 150^2 = 1440$$

$$s_y = \sqrt{1440} = 37{,}9473$$

$$s_{xy} = \sum_{j=1}^{p}\sum_{i=1}^{k} f_{ij}x_i y_j - \overline{\overline{xy}} = \frac{74790}{10} - 49{,}3.150 = 84$$

La ecuación de la recta de regresión de Y/X es:

$$y - \overline{y} = \frac{S_{xy}}{S_x^2}\left(x - \overline{x}\right) \text{ sustituyendo } y - 150 = \frac{84}{96{,}01}(x - 49{,}3)$$

Para un precio de 50€/Kg, la cantidad de producción esperada es:

$$y^*(50) = 150 + \frac{84}{96{,}01}(50 - 49{,}3) = 150{,}6124 Tm$$

Como medida de la bondad del ajuste calculamos el coeficiente de determinación:

$$R = r^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = \frac{84^2}{96{,}01.1440} = 0{,}051$$

b) $y^*(x+2) = 150 + \frac{84}{96{,}01}(x + 2 - 49{,}3) = y^*(x) + \frac{84}{96{,}01}2$

luego se incrementaría en $\frac{84}{96{,}01}2 = 1{,}75$

Por definición $R = \frac{S^2_{y^*}}{S_y^2} \Rightarrow S^2_{y^*} = RS_y^2 = 0{,}051.1440 = 73{,}44$.

Por otro lado

$$S_y^2 = S_e^2 + S^2_{y^*} \Rightarrow S_e^2 = S_y^2 - S^2_{y^*} = 1440 - 73{,}44 = 1366{,}56$$

Luego la parte de varianza no explicada por la regresión, que es la varianza residual es $S_e^2 = 1366{,}56$.